Improving the AUC of Probabilistic Estimation Trees

Cèsar Ferri[1], Peter Flach[2], José Hernández-Orallo[1]

{cferri, jorallo}@dsic.upv.es
Peter.Flach@bristol.ac.uk

[1]*Dep. de Sistemes Informàtics i Computació,Universitat Politècnica de València,València, Spain*
[2]*Department of Computer Science, University of Bristol, United Kingdom*

The 14th European Conference on Machine Learning
Cavtat, 22-26 September 2003

Introduction

- Many applications require some kind of reliability or numeric assessment of the quality of each classification.
- Soft classifiers can give an estimate of the reliability of each prediction
- Soft classifiers are useful in many scenarios, including combination of classifiers, cost-sensitive learning and safety-critical applications.

Probability Estimator Trees (1/3)

- A common presentation of a soft classifier is a probability estimator, i.e. a model that estimates for an example the probability of membership of every class.

- A decision tree adapted to be a probability estimator is called Probability Estimator Tree (PET).

Probability Estimator Trees (2/3)

- A trained decision tree can be easily adapted to be a PET by using the absolute class frequencies of each leaf of the tree.
  - If a node has the following absolute frequencies n_1, n_2, ..., n_c (obtained from the training dataset) the estimated probabilities for that node can be derived as $p_i = n_i / \Sigma n_i$.

- The probability estimates obtained by PETs are quite poor with respect to other probability estimators

Probability Estimator Trees (3/3)

- A good DTC is always a good PET??
  - There is a high correlation between quality of DTCs and quality of PETs, however many heuristics used for improving classification accuracy "reduce the quality of probability estimates" [Provost F., Domingos P., 2003]

It is worth investigating new heuristics and techniques which are specific to PETs

Evaluation of Probability Estimators

- The AUC (Area under the ROC Curve) measure has been a standard measure for evaluating the quality of PETs.

- We employ the Hand & Till extension of the AUC measure for multi-class problems.

Experimental Evaluation

- 50 datasets from the UCI repository (25 with 2 classes + 25 with more than 2 classes).
- Results show the average of 20x5-fold cross-validation (i.e. 100 executions for dataset).
- All experiments have been done within the SMILES system (http://www.dsic.upv.es/~flip/smiles/).
- GainRatio splitting criterion without node collapsing.

Smoothing

- We have investigated the effect of using probability smoothing in the leaves of PETs.
- Laplace smoothing $$p_i = \frac{n_i + 1}{\left(\sum_{i \in C} n_i\right) + c}$$
- m-estimate smoothing $$p_i = \frac{n_i + m \cdot p}{\left(\sum_{i \in C} n_i\right) + m}$$

Smoothing

- Results for smoothing:

	No smoothing	Laplace smoothing	M-estimate smoothing (m=4)
25 datasets (2 classes)	79.3	85.0	85.0
25 datasets (>2 classes)	78.0	83.9	84.0
All	78.7	84.5	84.5

m-branch smoothing (1/3)

- m-estimate and Laplace smoothing methods consider a uniform class distribution of the sample.
- The sample used to obtain the probability estimate in a leaf is the result of many sampling steps, as many as the depth of the leaf.
- It makes sense, then, to consider this history of samples when estimating the class probabilities in a leaf.

m-branch smoothing (2/3)

- Given a leaf node l and its associated branch $<n_1, n_2, \ldots, n_d>$ where $n_d = l$ and n_1 is the root, denote with n_{ij} the cardinality of class i at node n_j. Define $p_i^0 = 1/c$. We recursively compute the probabilities of the nodes from 1 to d as follows:

$$p_i^j = \frac{n_i^j + m \cdot p_i^{j-1}}{\left(\sum_{i \in C} n_i^j\right) + m}$$

- The m-branch smoothed probabilities of leaf l are given by p_i^d.

m-branch smoothing (3/3)

- Results for m-branch smoothing:

	M-estimate smoothing (m=4)	m-branch smoothing (m=4)	Better?
25 datasets (2 classes)	85.0	85.8	8 ✓ 14 = 3 ×
25 datasets (>2 classes)	90.5	91.5	13 ✓ 9 = 3 ×
All	87.8	88.7	21 ✓ 23 = 6 ×

Splitting Criteria for PETs

- The splitting criterion is a crucial factor in the learning of decision trees
- Classical splitting criteria (Gini, Gain Ratio, DKM) have been designed and evaluated for classifiers, not for probability estimators.

MAUC splitting criterion

- In previous works we have introduced a novel criterion aimed at maximising the AUC of the resulting tree rather than its accuracy.
- It simply computes the quality of each split as the AUC of the nodes resulting from that split, assuming a two-class problem.
- The splitting criterion can be generalised for multiclass problems by using the Hand and Till's 1-vs-1 average.

MSE splitting criterion

- A different approach is to consider that the tree really predicts probabilities.
- It makes sense to minimise the quadratic error (MSE) committed when guessing these probabilities.
- Given a split *s*, the quality of the split is defined as:

$$MSEEsplit(s) = \sum_{k=1..n} q_k \cdot \left(- \sum_{i=1..c} Error_i \right)$$

Splitting criteria comparison

- Results in AUC measure:

	C4.5	Gain	Mgini	DKM	MAUCSplit	MSESplit	MSESplit Vs. C4.5
25 datasets (2 classes)	84.9	84.8	84.6	84.8	85.0	85.3	7 ✓ 13 = 5 ×
25 datasets (>2 classes)	90.6	90.9	90.8	91.1	90.8	90.9	7 ✓ 13 = 5 ×
All	87.7	87.8	87.7	87.9	87.8	88.1	14 ✓ 26 = 10 ×

Pruning and PETs

- Some works have shown that pruning is counterproductive for obtaining good PETs.
- The better the smoothing at the leaves is the worse pruning will be.
- It is interesting to design pruning methods that reduce the size of the tree without degrading too much the quality of the PET.

Cardinality-based pruning

- The size of the sample is crucial to establish the quality of a probability estimation.

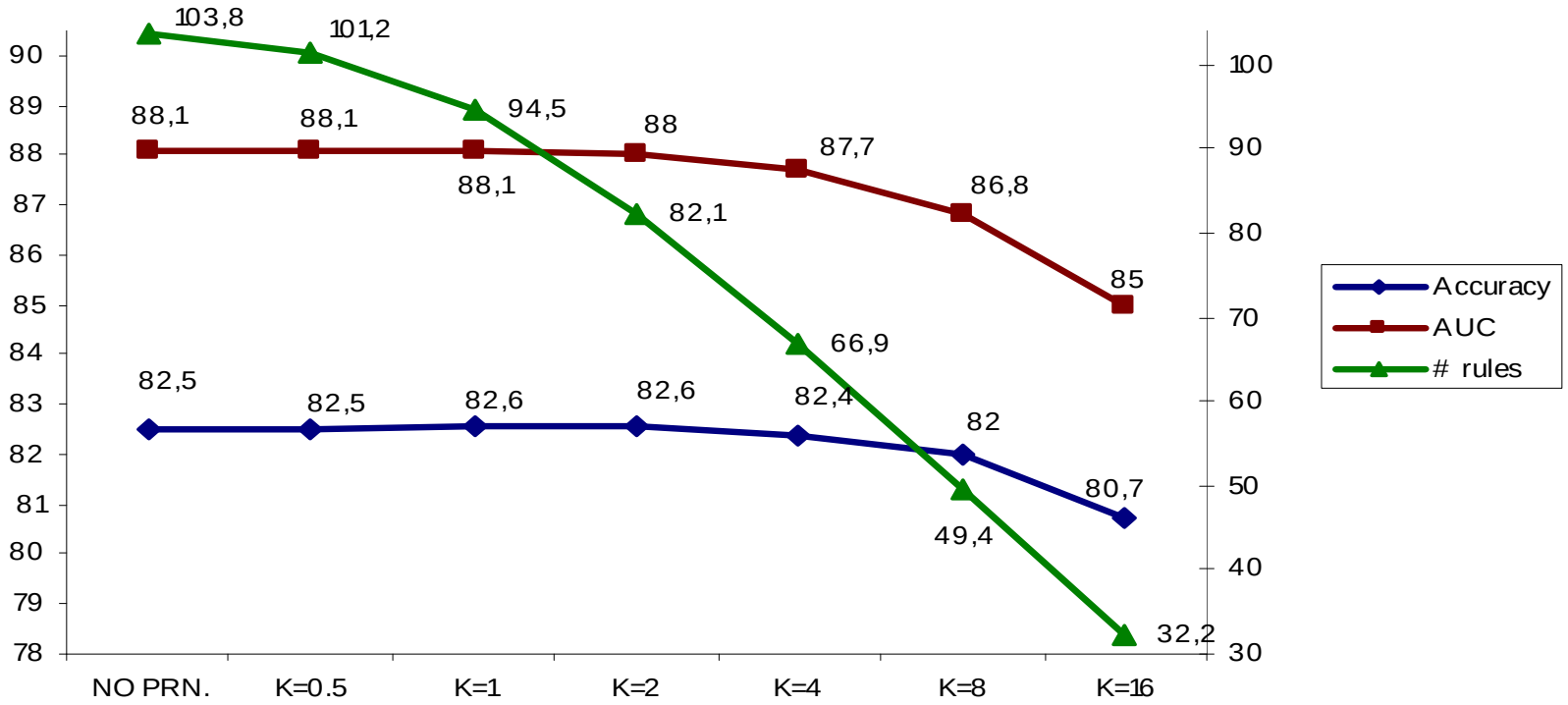
- The poorest estimates of a PET will be obtained by the smallest nodes. It is also important to consider the number of classes.

- Given a node *l*, it will be pruned when:

$$Card(l) < 2\frac{K}{c}$$

where *Card(l)* is the cardinality of node *l*, *K* is a constant (*K*=0 means no pruning) and *c* is the number of classes.

Cardinality-based pruning



Summary

	C4.5	C4.5 with Laplace Smoothing	C4.5 with m-branch Smoothing	MSESplit with m-branch Smoothing	C4.5 + laplace vs MSESplit + m-branch + pruning k=1
25 datasets (2 classes)	78.0	83.9	84.9	85.4	11 ✓ 9 = 5 ×
25 datasets (>2 classes)	87.3	89.5	90.6	90.6	16 ✓ 4 = 5 ×
All	82.5	86.7	87.7	88.1	27 ✓ 13 = 10 ×

Conclusions (1/2)

- We have reassessed the construction of PETs, evaluating and introducing new methods:
  - A new smoothing correction that takes the whole branch of decisions into account.
  - A novel MSEE splitting criterion aimed at reducing the squared error of the probability estimate.
  - A simple cardinality pruning method can be applied to obtain simpler PETs without degrading their quality too much.

Conclusions (2/2)

- An exhaustive experimental evaluation has shown the performance of the methods
- One of the first works that that compares the ranking of probability estimates of several splitting criteria for PETs
- **SMILES** is freely available at:
  - **http://www.dsic.upv.es/~flip/smiles/**

Future work

- The study of methods for improving the estimates without modifying the structure of a single tree.
- The design of better pruning methods for PETs.
- The use of PET's in ensemble methods.